

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN



Lab 01 - Preprocessing

Data Mining - Term I/2020-2021

Sinh viên thực hiện:

1. Nguyễn Thị Thu Hằng - 18120027
2. Phạm Thị Hoài Hiền - 18120178

Thành phố Hồ Chí Minh, ngày 2 tháng 11 năm 2020

I. Đánh giá mức độ hoàn thành:

MSSV	Họ và tên	Tỉ lệ thực hiện
18120027	Nguyễn Thị Thu Hằng	50%
18120178	Phạm Thị Hoài Hiền	50%

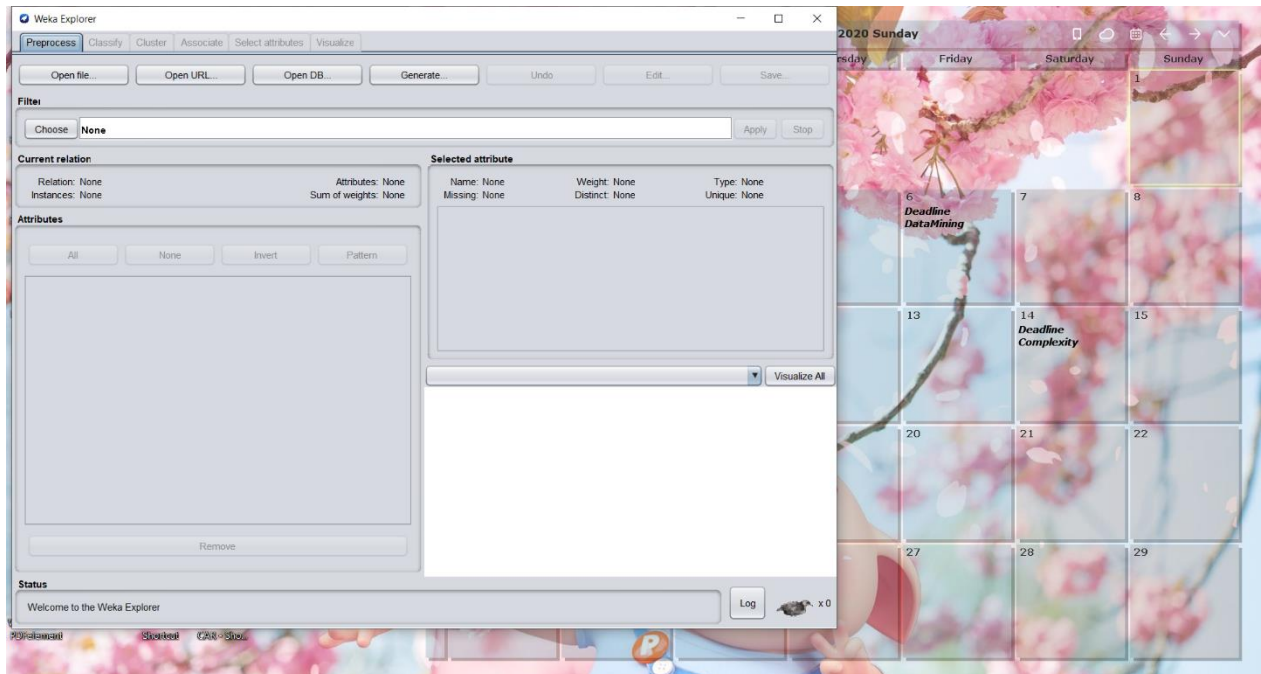
Mức độ hoàn thành chung: 100%

II. Yêu cầu 1: Cài đặt Weka

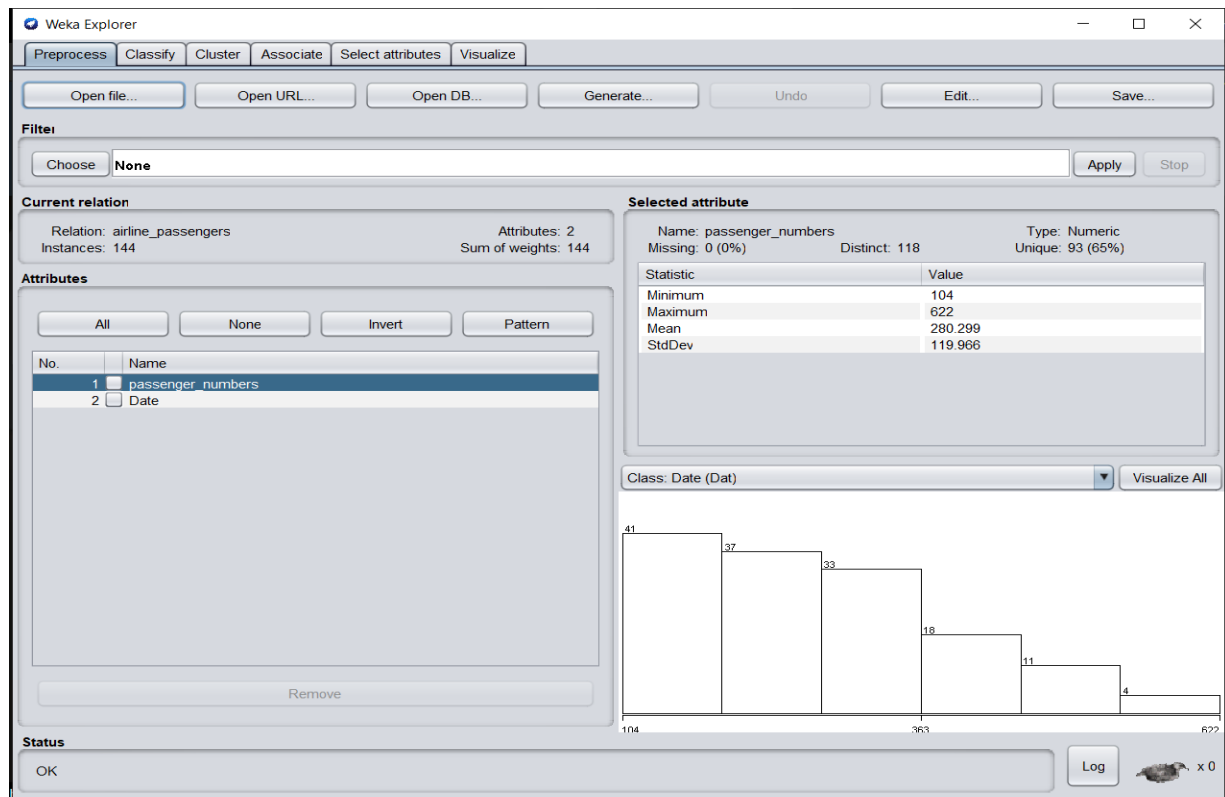
1. *Mức độ hoàn thành:* 100%

2. *Chi tiết:*

Giao diện Explorer sau khi cài đặt xong Weka:

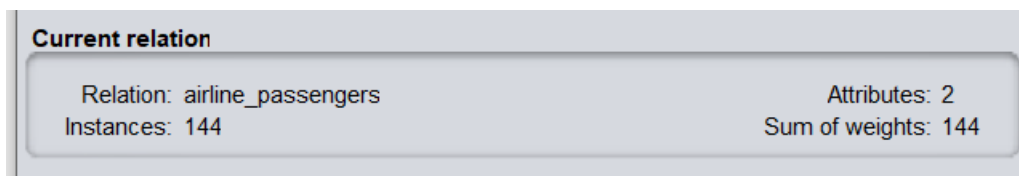


Chọn mở tệp *airline.arff*, ta có hình ảnh như sau:

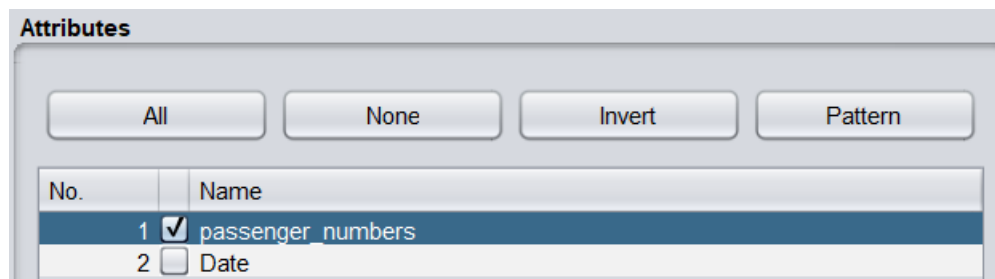


Giải thích ý nghĩa:

- *Nhóm điều khiển Current Relation:* thể hiện tên cơ sở dữ liệu đang được tải. Tên thể hiện (Relation) là *airline_passengers*; số mẫu (Instances) là 144; số lượng thuộc tính của bảng (Attributes) là 2

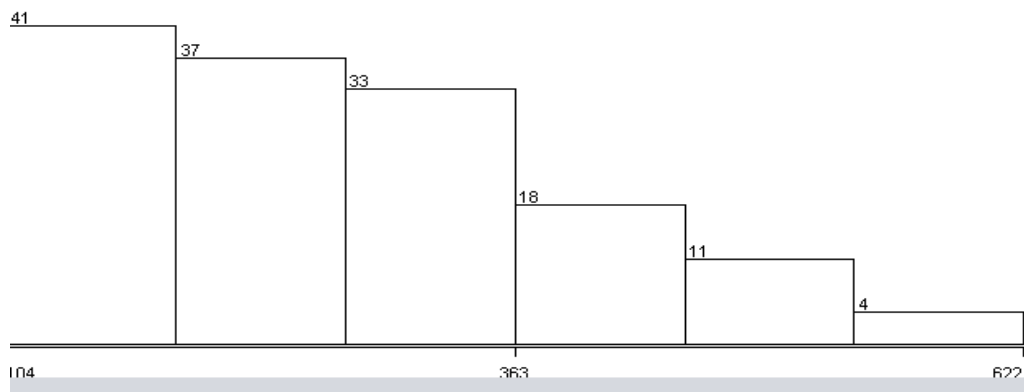


- *Nhóm điều khiển Attributes:* hiển thị các thuộc tính (attribute) trong cơ sở dữ liệu. Khi ta chọn một thuộc tính từ danh sách này thì chi tiết thuộc tính được hiển hiện trong **Selected Attributes**



- *Nhóm điều khiển Selected Attribute:* khi ta chọn thuộc tính tại **Attributes** thì nó sẽ hiện ra một số thông tin của thuộc tính được chọn:
 - + Tên và loại thuộc tính được hiển thị
 - + Số lượng giá trị phân biệt (Distinct)

Selected attribute	
Name: passenger_numbers	Type: Numeric
Missing: 0 (0%)	Distinct: 118
	Unique: 93 (65%)
Statistic	Value
Minimum	104
Maximum	622
Mean	280.299
StdDev	119.966
Class: Date (Dat)	
Visualize All	



Giải thích các tab trong Explorer:

- *Preprocess Tab:* Ban đầu khi bạn mở trình khám phá, chỉ tab Tiền xử lý được bật. Bước đầu tiên trong học máy là xử lý trước dữ liệu. Do đó, trong tùy chọn Preprocess, bạn sẽ chọn tệp dữ liệu, xử lý và làm cho nó phù hợp để áp dụng các thuật toán học máy khác nhau.
- *Classify Tab:* Cung cấp cho bạn một số thuật toán máy học cho việc phân loại dữ liệu của bạn. Để liệt kê một số, bạn có thể áp dụng các

thuật toán như Hồi quy tuyến tính, Hồi quy logistic, Máy vector hỗ trợ, Cây quyết định, RandomTree, RandomForest, NaiveBayes, v.v. Danh sách này rất đầy đủ và cung cấp cả thuật toán học máy có giám sát và không giám sát.

- *Cluster Tab*: Cung cấp một số thuật toán phân cụm được cung cấp - chẳng hạn như SimpleKMeans, FilteredClusterer, HierarchicalClusterer, v.v.
- *Associate Tab*: Để khám phá các luật kết hợp từ dữ liệu, chứa các thuật toán Apriori, FilteredAssociator và FPGrowth.
- *Select Attributes Tab*: Cho phép bạn làm nổi bật các thuộc tính liên quan (quan trọng nhất) của dữ liệu, dựa trên một số thuật toán như ClassifierSubsetEval, PrincipalComponents, v.v.
- *Visualize Tab*: Trực quan hóa dữ liệu đã xử lý của mình để phân tích.

III. Yêu cầu 2:

1. **Mức độ hoàn thành:** 100%

2. **Khám phá tập dữ liệu Ung thư vú (breast_cancer.arff)**

a. Tập dữ liệu có bao nhiêu mẫu (instances)?

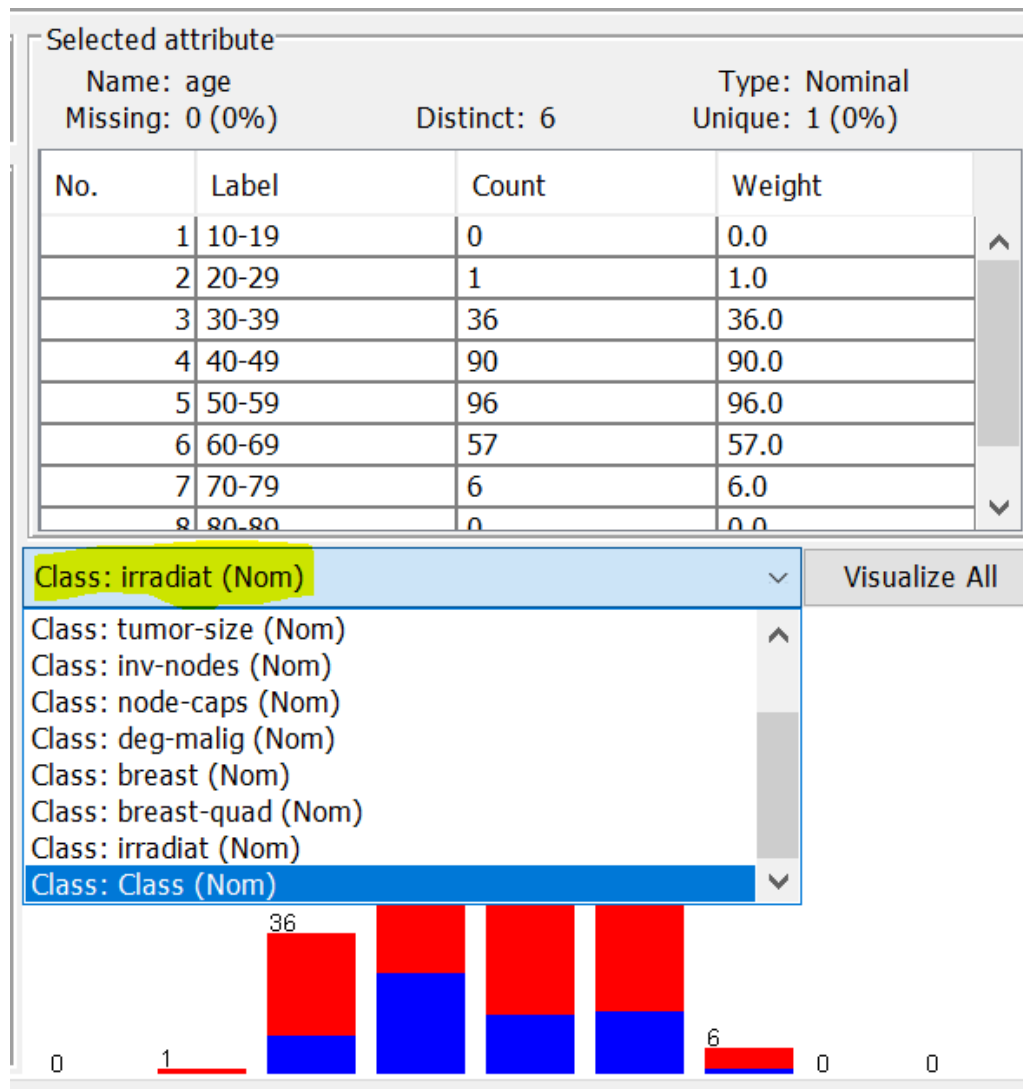
Trả lời: 286

b. Tập dữ liệu có bao nhiêu thuộc tính (attributes)?

Trả lời: 10

c. Thuộc tính nào được dùng làm lớp (class)? Có thể thay đổi thuộc tính dùng làm lớp hay không? Nếu có thì bằng cách nào?

Trả lời: thuộc tính dùng làm lớp là *class*, có thể thay đổi thuộc tính dùng làm lớp bằng cách chọn trong tab class như hình dưới:



d. Tìm hiểu chi tiết từng thuộc tính trong khung Attributes và cho biết: có bao nhiêu thuộc tính bị thiếu dữ liệu (missing values)? Thuộc tính nào thiếu dữ liệu ít nhất/nhiều nhất? Trình bày tổng quát các cách để giải quyết vấn đề missing values.

Trả lời: có 2 thuộc tính bị thiếu dữ liệu là *node-caps* và *breast-quad*. Thuộc tính bị thiếu dữ liệu nhiều nhất là *node-caps*, ít nhất là *breast-quad*

Selected attribute		
Name: node-caps	Type: Nominal	
Missing: 8 (3%)	Distinct: 2	Unique: 0 (0%)

- Selected attribute		
Name: breast-quad	Type: Nominal	
Missing: 1 (0%)	Distinct: 5	Unique: 0 (0%)

Cách để giải quyết vấn đề missing value trong Weka: chọn thuộc tính cần thực hiện: trong mục *filter*, ta chọn mục *unsupervised* -> *attribute* -> *ReplaceMissingValues*, sau đó nhấn *Apply* để áp dụng. Ở đây giá trị bị thiếu sẽ được thay bằng mode (giá trị trung bình) và mean (giá trị xuất hiện thường xuyên nhất) từ tập huấn luyện.

Filter					
Choose	ReplaceMissingValues				Apply Stop
Current relation			- Selected attribute		
Relation: breast-cancer-wek...	Attributes: 10		Name: node-caps	Type: Nominal	
Instances: 286	Sum of weights: 286		Missing: 0 (0%)	Distinct: 2	Unique: 0 (0%)

- e. *Giải thích ý nghĩa của đồ thị trong cửa sổ Explorer. Bạn đặt tên cho đồ thị này là gì? Màu xanh và màu đỏ có nghĩa gì? Đồ thị này biểu diễn cho cái gì?*

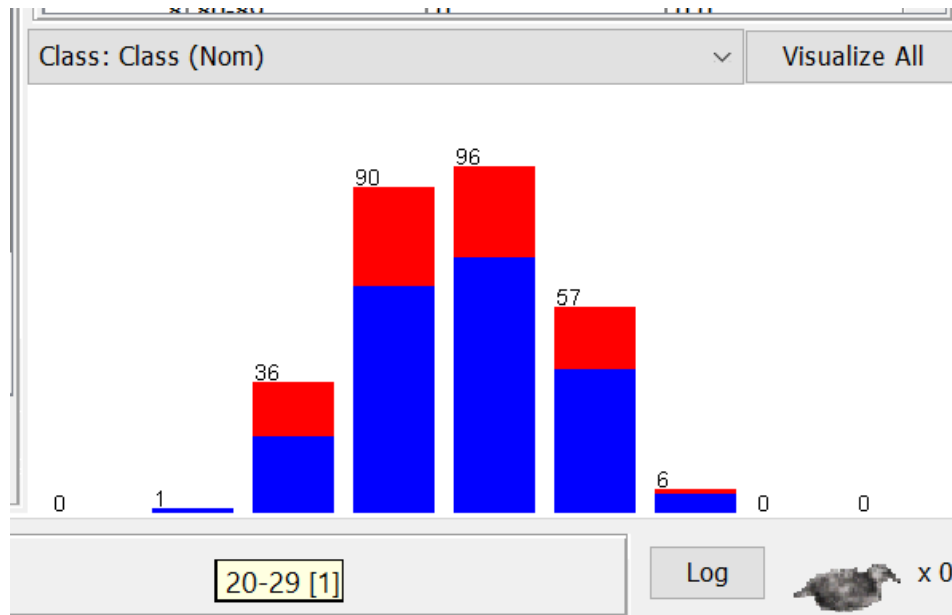
Trả lời: Đồ thị trong cửa sổ Explorer là đồ thị của thuộc tính ta đang chọn, được tô màu theo thuộc tính chọn làm lớp (chỉ có những thuộc tính định danh (nominal) mới có thể tô màu). Màu xanh và màu đỏ thể hiện số lượng giá trị của thuộc tính lớp trong thuộc tính đang xét. Thuộc tính lớp có bao nhiêu giá trị thì có bấy nhiêu màu.

Ví dụ, thuộc tính lớp là *class*, thuộc tính chọn để xét là *age*. Màu xanh là *no-recurrence-events*, màu đỏ là *recurrence-events*. Ta thấy từ độ tuổi 20-29 chỉ có một giá trị là *no-recurrence-events* nên cột 20-29 toàn màu xanh.

Viewer

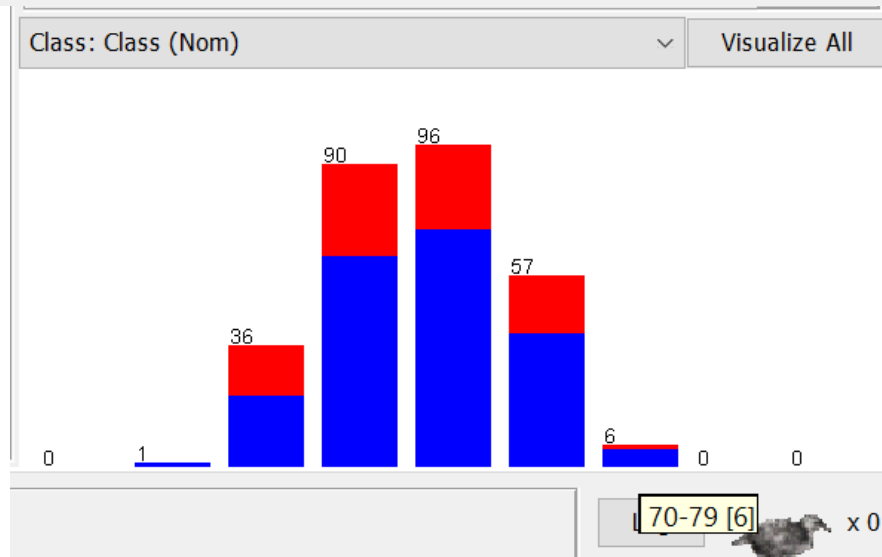
Relation: breast-cancer

No.	1: age Nominal	2: menopause Nominal	3: tumor-size Nominal	4: inv-nodes Nominal	5: node-caps Nominal	6: deg-malig Nominal	7: breast Nominal	8: breast-quad Nominal	9: irradiat Nominal	10: Class Nominal
1	20-29	premeno	35-39	0-2	no	2	right	right_up	no	no-recurrence-events
2	30-39	premeno	20-24	0-2	no	3	left	central	no	no-recurrence-events
3	30-39	premeno	15-19	6-8	yes	3	left	left_low	yes	recurrence-events



Còn độ tuổi 70-79 chỉ có một *recurrence-events* nên có màu đỏ ở cột đó.

281	70-79	ge40	15-19	9-11		1	left	left_low	yes	recurrence-events
282	70-79	ge40	40-44	0-2	no	1	right	right_up	no	no-recurrence-events
283	70-79	ge40	40-44	0-2	no	1	right	left_up	no	no-recurrence-events
284	70-79	ge40	10-14	0-2	no	2	left	central	no	no-recurrence-events
285	50-59	ge40	0-4	0-2	no	1	left	right_low	no	no-recurrence-events
286	70-79	ge40	20-24	0-2	no	3	left	left_up	no	no-recurrence-events



3. Khám phá tập dữ liệu Weather

- Tập dữ liệu có bao nhiêu thuộc tính? Bao nhiêu mẫu? Phân loại các thuộc tính theo kiểu dữ liệu (categorical/numeric). Thuộc tính nào là lớp?

Trả lời:

- Tập dữ liệu có 5 thuộc tính, 14 mẫu.
- Phân loại thuộc tính theo kiểu dữ liệu: **Nominal** (thuộc tính *outlook*, *windy*, *play*), **Numeric** (thuộc tính *temperature*, *humidity*)
- Thuộc tính lớp là thuộc tính *play*

Selected attribute			
Name: outlook		Type: Nominal	
Missing: 0 (0%)		Distinct: 3	
		Unique: 0 (0%)	
No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	4	4.0
3	rainy	5	5.0

b. Liệt kê five-number summary của thuộc tính temperature và humidity. Weka có cung cấp những giá trị này không?

Trả lời: *five-number summary* là một tập thống kê mô tả cung cấp thông tin cho một dataset. Nó bao gồm mẫu nhỏ nhất (minimum), mẫu lớn nhất (maximum), trung vị (median), giá trị phân vị thứ nhất (first quartile), giá trị phân vị thứ ba (third quartile). (*theo wikipedia*). Trong Weka đối với thuộc tính có kiểu dữ liệu số, Weka cung cấp 4 thông tin: giá trị lớn nhất, nhỏ nhất, giá trị trung bình và độ lệch chuẩn

- Thuộc tính *temperature*:

Selected attribute	
Name: temperature	
Missing: 0 (0%)	Distinct: 12
Type: Numeric	
Unique: 10 (71%)	
Statistic	Value
Minimum	64
Maximum	85
Mean	73.571
StdDev	6.572

- Thuộc tính *humidity*:

Selected attribute	
Name: humidity	
Missing: 0 (0%)	Distinct: 10
Type: Numeric	
Unique: 7 (50%)	
Statistic	Value
Minimum	65
Maximum	96
Mean	81.643
StdDev	10.285

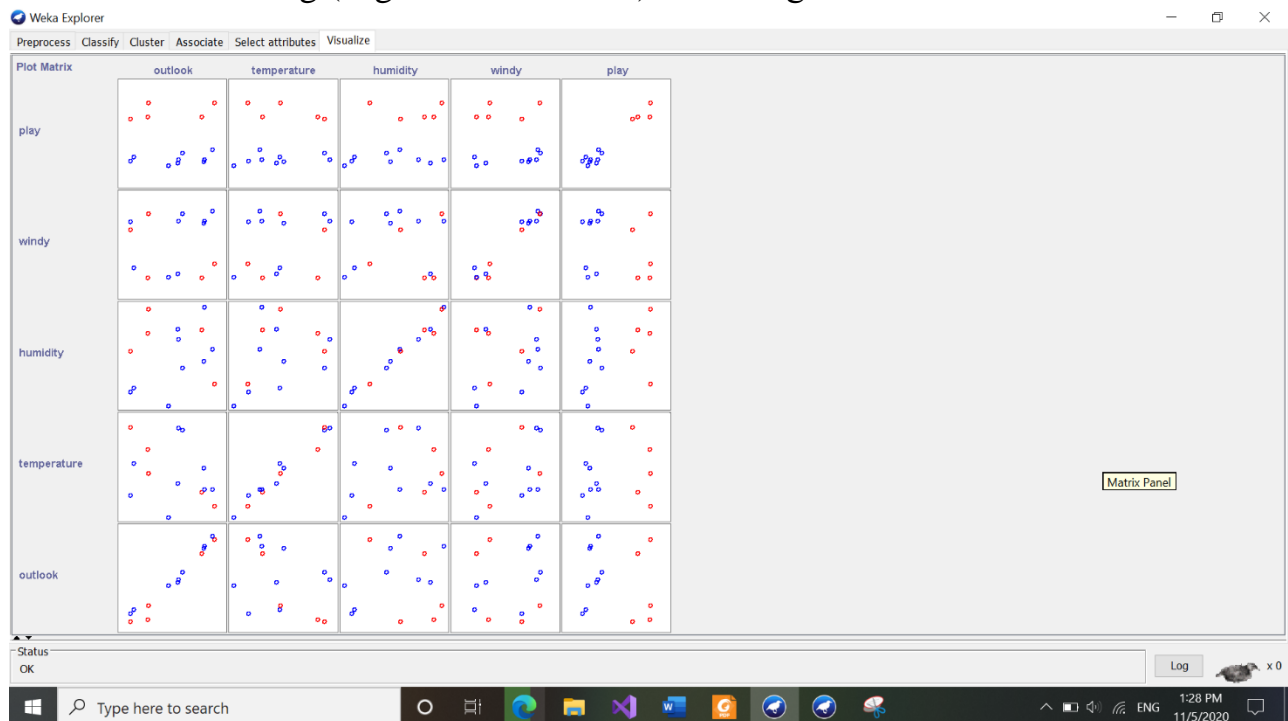
c. Lần lượt xem xét các thuộc tính khác của dataset dưới dạng đồ thị.
Dán các ảnh chụp màn hình vào bài làm

Trả lời: chọn mục *Visualize all* để xem tất cả đồ thị của các thuộc tính



d. Chuyển sang tab Visualize. Thuật ngữ sử dụng trong textbook để đặt tên cho các đồ thị ở đây là gì? Chọn jitter tối đa để thấy tổng quan hơn về phân bố dữ liệu. Theo bạn có những cặp thuộc tính khác nhau nào có vẻ như tương quan với nhau không?

Trả lời: Thuật ngữ sử dụng trong textbook để đặt tên cho các đồ thị là **scatter-plot matrix (ma trận biểu đồ phân tán)**. Không có cặp thuộc tính khác nhau nào có vẻ tương quan với nhau, vì biểu đồ phân tán của nó không có dạng hướng lên (positive correlation) hay hướng xuống (negative correlation) như trong textbook mô tả.



4. Khám phá tập dữ liệu Tín dụng Đức

a. Nội dung của phần ghi chú (comment) trong *credit-g.arff* (khi mở bằng 1 text editor bất kì) nói về điều gì? Tập dữ liệu có bao nhiêu mẫu? Bao nhiêu thuộc tính? Mô tả 5 thuộc tính bất kì (phải vừa có cả thuộc tính rời rạc và thuộc tính liên tục).

Trả lời:

- Nội dung ghi chú nói về mô tả của dataset, gồm có tên, nguồn thông tin, số lượng mẫu, số lượng thuộc tính, mô tả thuộc tính. Có hai dataset được cung cấp, bản gốc của giáo sư Hofmann, chứa các thuộc tính phân loại/kí hiệu nằm trong tệp

german.data. Ghi chú: “đối với thuật toán cần thuộc tính số (*numeric*), Đại học Strathclyde tạo ra tệp *greman.data-numeric*, tệp này đã được chỉnh sửa cho phù hợp với thuật toán không thể dùng các biến phân loại (*category*). Một số thuộc tính phân loại được mã hoá dưới dạng số nguyên”.

- Tập dữ liệu có 1000 mẫu, 21 thuộc tính
- Mô tả 5 thuộc tính bất kì:
 - + *checking_status* (kiểu dữ liệu *nominal*): trạng thái của tài khoản hiện có, xét lương ít nhất 1 lần/năm (gồm các mức như sau: <0 ; $0 \leq X \leq 200$; ≥ 200 ; no checking)
 - + *duration* (kiểu dữ liệu *numeric*): thời gian sử dụng tài khoản (tính theo tháng). Trong tập dữ liệu thì thời gian nhỏ nhất là 4, lớn nhất là 72, trung bình là 20,093
 - + *credit_history* (kiểu dữ liệu *nominal*): lịch sử tín dụng, gồm các nhãn *no credits*,
 - + *credit_amount* (kiểu dữ liệu *nominal*): số dư trong thẻ, thấp nhất là 250 và cao nhất là 18424, trung bình là 3271.258
 - + *employment* (kiểu dữ liệu *nominal*): số năm làm việc, gồm có unemployed (đang thất nghiệp), <1 ; $1 \leq X < 4$; $4 \leq X < 7$; ≥ 7

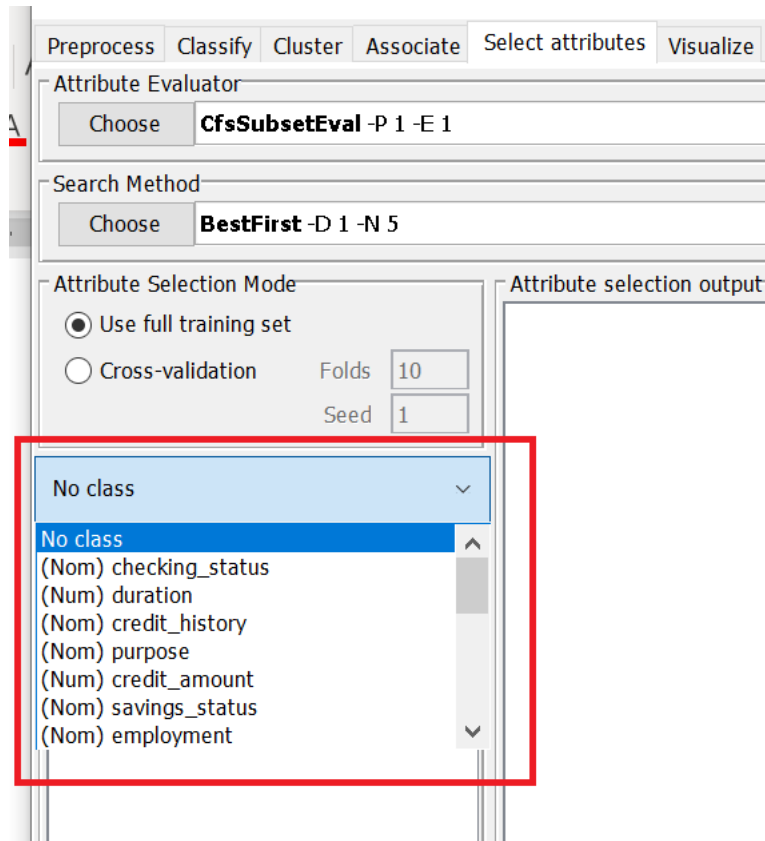
b. Tên của thuộc tính lớp là gì? Đánh giá phân bố của các lớp, tức là cân bằng hay lệch về một lớp?

Trả lời: tên thuộc tính lớp là *class*. Phân bố của các lớp đa phần bị lệch, chỉ có các lớp sau là cân bằng: *credit_history*, *employment*, *personal_status*, *property_magnitude*, *housing*, *job*, *own_telephone*.

c. Sử dụng tab *Select attributes*. Liệt kê những lựa chọn khác nhau của Weka để chọn lọc thuộc tính, giải thích ngắn gọn từng phương pháp.

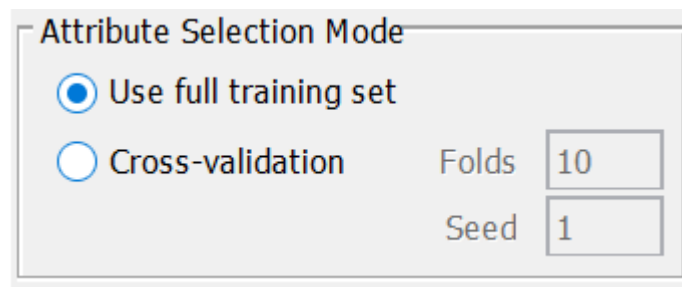
Trả lời: để lựa chọn thuộc tính, ta cần thiết lập bốn đối tượng cụ thể sau:

- *Lựa chọn thuộc tính được dự đoán (biến phụ thuộc):* Sử dụng dropdown liệt kê tập thuộc tính của tập dữ liệu



- **Bộ đánh giá thuộc tính (Attribute Evaluator):** Để đánh giá tập các thuộc tính của tập dữ liệu. WEKA cung cấp 11 phương pháp đánh giá thuộc tính, gồm:
 - **CfsSubsetEval:** Đánh giá tập thuộc tính bằng cách xem xét khả năng dự đoán của từng thuộc tính riêng lẻ và mức độ dư thừa giữa chúng.
 - **ClassifierSubsetEval:** Đánh giá tập thuộc tính con trong tập huấn luyện (training) hoặc tập kiểm tra (test) riêng biệt.
 - **ClassifierAttributeEval:** Đánh giá thuộc tính bằng cách sử dụng bộ phân lớp do người dùng chọn.
 - **CorrelationAttributeEval:** Đánh giá một thuộc tính dựa trên sự tương quan với lớp.
 - **GainRatioAttributeEval:** Đánh giá một thuộc tính dựa trên tỷ lệ gia tăng.
 - **InfoGainAttributeEval:** Đánh giá một thuộc tính dựa trên thông tin thu được.
 - **OneRAttributeEval:** Đánh giá một thuộc tính bằng cách sử dụng bộ phân loại OneR.

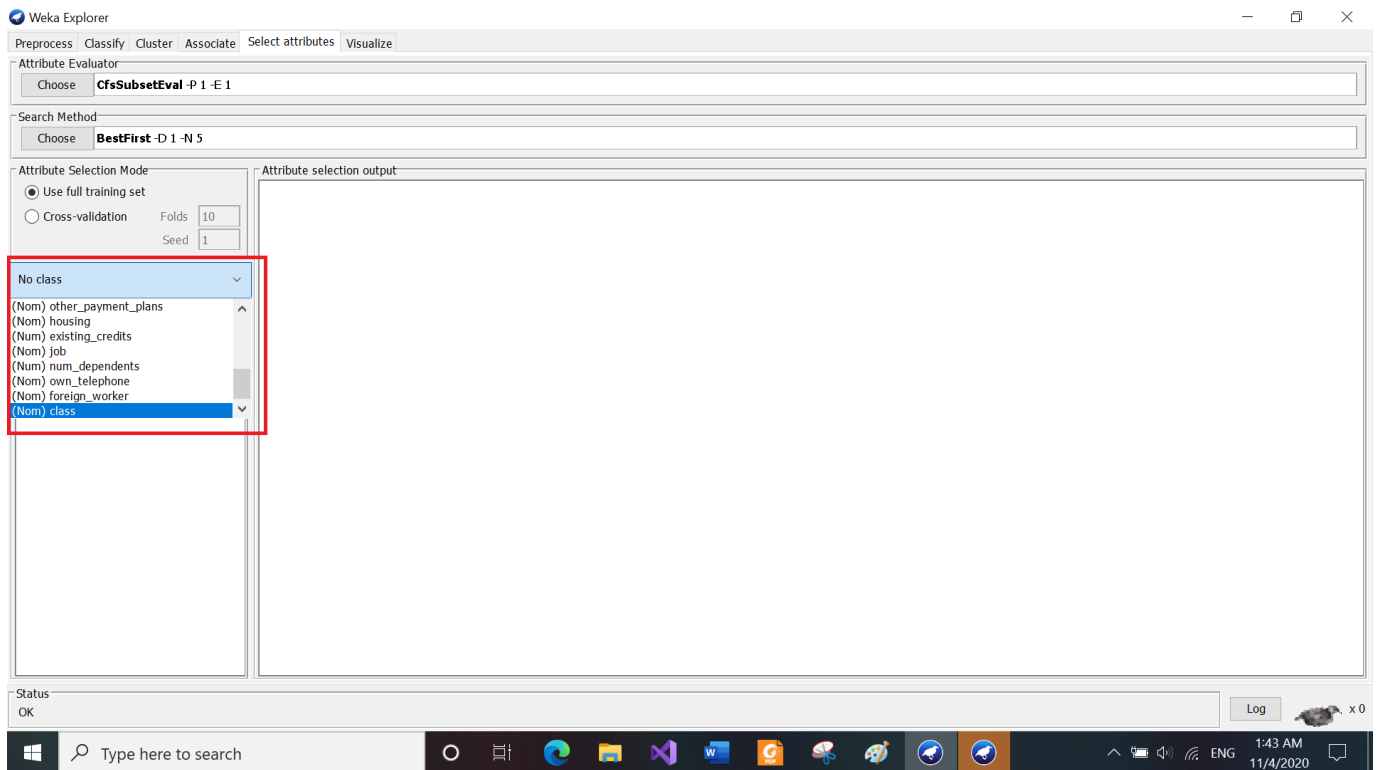
- **PrincipalComponents:** Thực hiện phân tích thành phần chính và chuyển đổi dữ liệu.
- **ReliefFAttributeEval:** Đánh giá thuộc tính dựa trên các thể hiện.
- **SymmetricalUncertAttributeEval:** Đánh giá một thuộc tính dựa trên bất đối xứng.
- **WrapperSubsetEval:** Đánh giá tập thuộc tính dựa trên một bộ phân loại cùng với xác nhận chéo.
- **Phương thức tìm kiếm (Search Method):** Để xác định phương pháp tìm kiếm được thực hiện. WEKA cung cấp 3 phương thức tìm kiếm, gồm:
 - **BestFirst:** Tiến hành kỹ thuật leo đồi tham lam kết hợp với quay lui.
 - **GreedyStepwise:** Thực hiện tìm kiếm tham lam về phía trước hoặc phía sau thông qua không gian các tập con thuộc tính.
 - **Ranker:** Xếp hạng các thuộc tính theo đánh giá trọng số của từng thuộc tính. Sử dụng kết hợp với các bộ đánh giá thuộc tính (ReliefF, GainRatio,...).
- **Chế độ lựa chọn thuộc tính (Attribute Selection Mode):** Xác định chế độ lựa chọn thuộc tính, sử dụng tập huấn luyện đầy đủ hoặc tiến hành xác nhận chéo. Để xây dựng mô hình hồi quy tuyến tính, cần lựa chọn sử dụng tập huấn luyện đầy đủ.



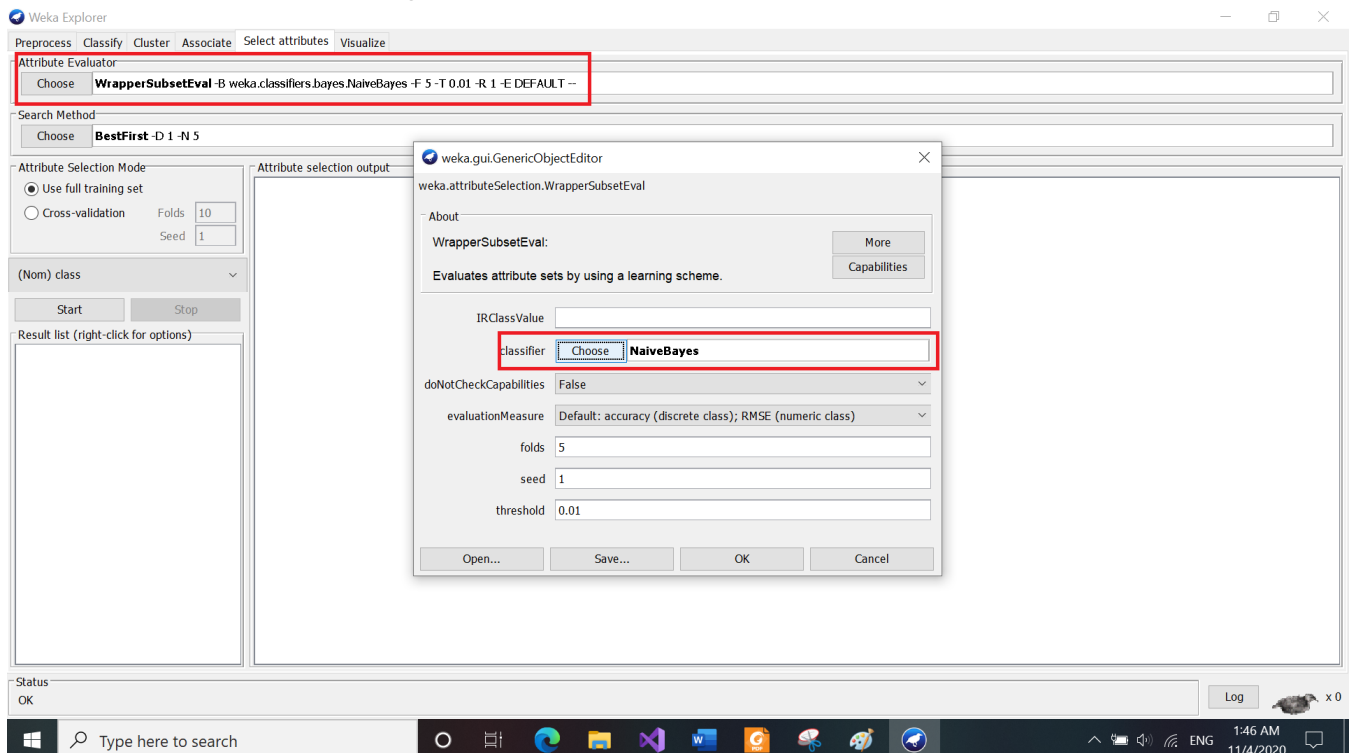
d. Cần sử dụng bộ lọc nào để chọn ra 5 thuộc tính có tương quan cao nhất với thuộc tính lớp? Mô tả các bước làm, kèm theo hình chụp từng bước và kết quả cuối cùng

Trả lời:

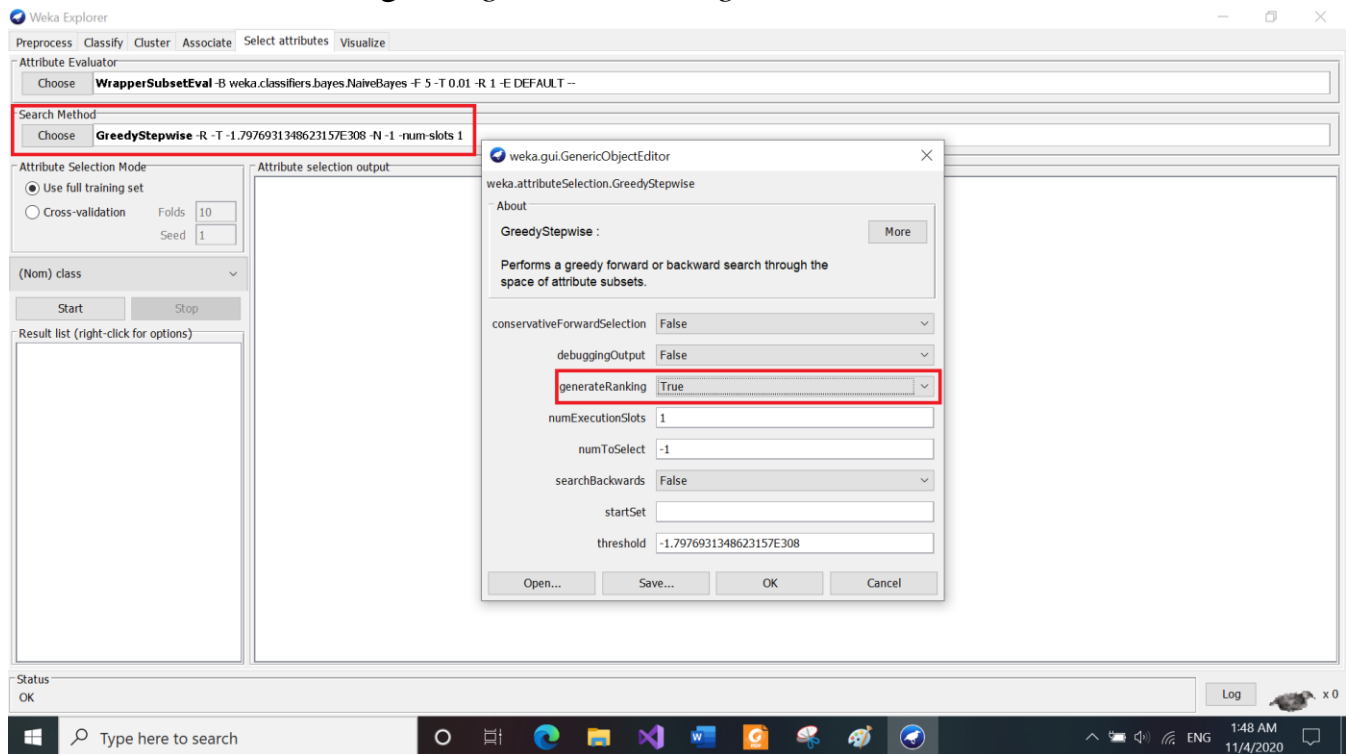
- Bước 1: Qua tab *Selected Attribute*, chọn *class* là lớp



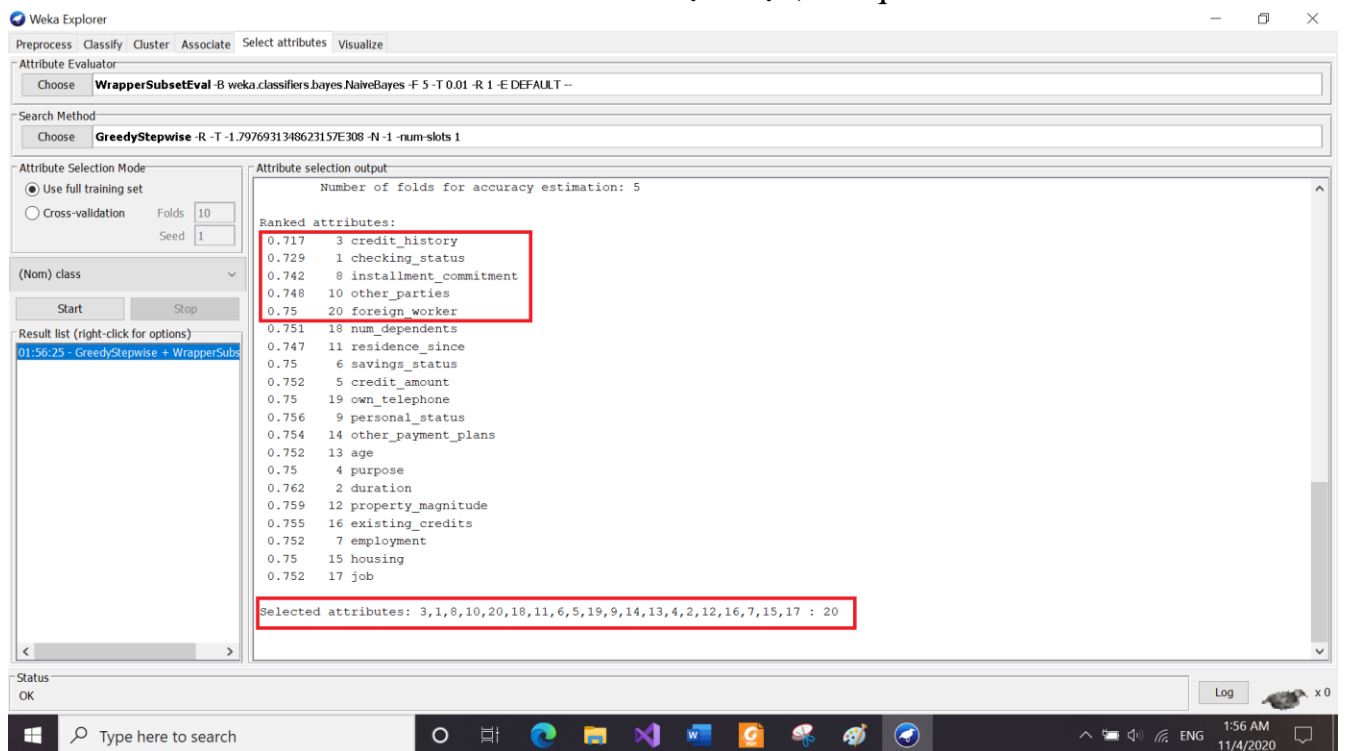
- Bước 2: Trong *Attribute Evaluator* chọn *WrapperSubsetEval*. Trong mục *classifier* chọn *NaiveBayes*



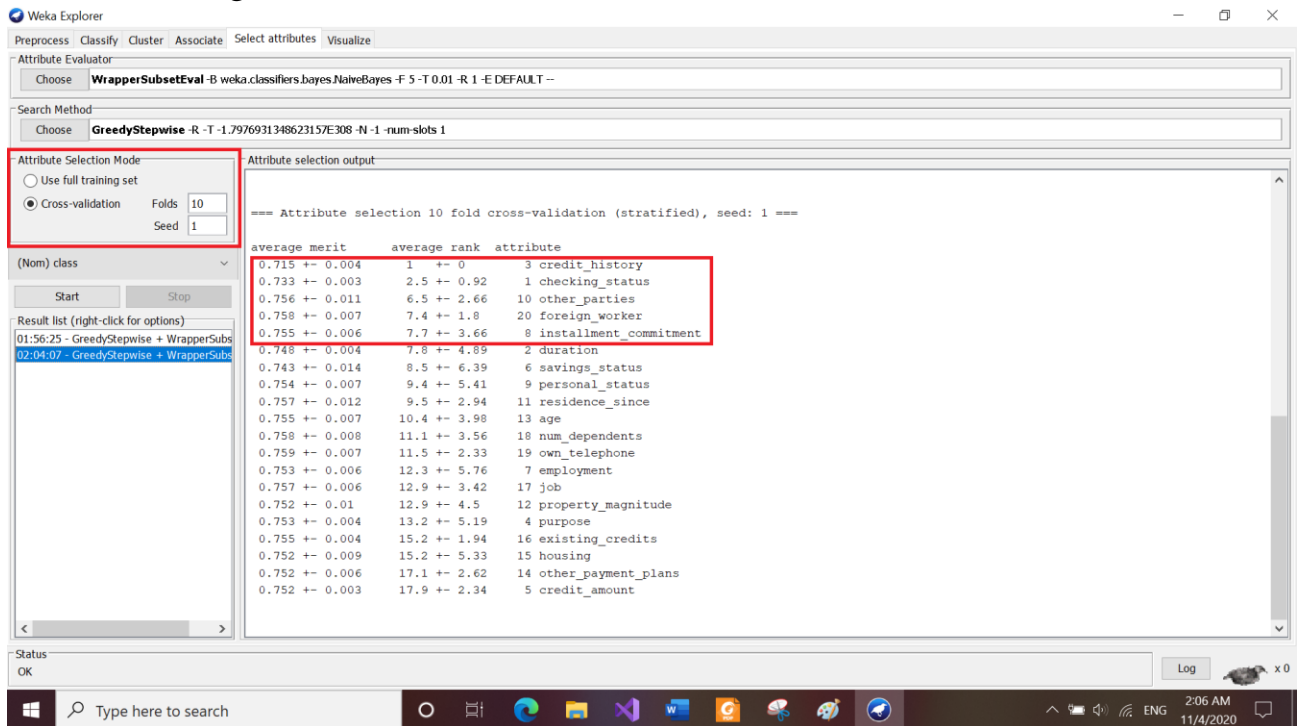
- Bước 3: Trong mục *Search Method* chọn *GreedyStepwise*, trong mục *generateRanking* chọn *True*:



- Bước 4: Bấm *Start* để thực hiện, kết quả như sau:



Vậy 5 thuộc tính có tương quan cao nhất với thuộc tính lớp là *credit_history*, *checking_status*, *installment_commitment*, *other_parties*, *foreign_worker*. Đổi chế độ chọn thuộc tính sang *cross-validation*, ta vẫn được kết quả tương tự, chỉ khác thứ tự xếp hạng:



Attribute Selection Mode:

- ☐ Use full training set
- ☒ Cross-validation

Folds: 10
Seed: 1

Attribute selection output:

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===
```

average merit	average rank	attribute
0.715 +/- 0.004	1 +/- 0	3 credit_history
0.733 +/- 0.003	2.5 +/- 0.92	1 checking_status
0.756 +/- 0.011	6.5 +/- 2.66	10 other_parties
0.758 +/- 0.007	7.4 +/- 1.8	20 foreign_worker
0.755 +/- 0.006	7.7 +/- 3.66	8 installment_commitment
0.748 +/- 0.004	7.8 +/- 4.89	2 duration
0.743 +/- 0.014	8.5 +/- 6.39	6 savings_status
0.754 +/- 0.007	9.4 +/- 5.41	9 personal_status
0.757 +/- 0.012	9.5 +/- 2.94	11 residence_since
0.755 +/- 0.007	10.4 +/- 3.98	13 age
0.758 +/- 0.008	11.1 +/- 3.56	18 num_dependents
0.759 +/- 0.007	11.5 +/- 2.33	19 own_telephone
0.753 +/- 0.006	12.3 +/- 5.76	7 employment
0.757 +/- 0.006	12.9 +/- 3.42	17 job
0.752 +/- 0.01	12.9 +/- 4.5	12 property_magnitude
0.753 +/- 0.004	13.2 +/- 5.19	4 purpose
0.755 +/- 0.004	15.2 +/- 1.94	16 existing_credits
0.752 +/- 0.009	15.2 +/- 5.33	15 housing
0.752 +/- 0.006	17.1 +/- 2.62	14 other_payment_plans
0.752 +/- 0.003	17.9 +/- 2.34	5 credit_amount

IV. Yêu cầu 3:

1. **Mức độ hoàn thành:** 100%
2. **Chi tiết:** xem trong source code